SIAM/ASA J. Uncertainty Quantification
Vol. 9, No. 1, pp. 135–162

Using Coupling Methods to Estimate Sample Quality of Stochastic Differential Equations[*]

Matthew Dobson[†], Yao Li[†], and Jiayu Zhai[†]

Abstract. A probabilistic approach to estimating sample qualities of stochastic differential equations is introduced in this paper. The aim is to provide a quantitative upper bound of the distance between the invariant probability measure of a stochastic differential equation and that of its numerical approximation. In order to extend estimates of finite time truncation error to infinite time, it is crucial to know the rate of contraction of the transition kernel of the SDE. We find that suitable numerical coupling methods can effectively estimate such rate of contraction, which gives the distance between two invariant probability measures. Our algorithms are tested with several low and high dimensional numerical examples.

Key words. stochastic differential equation, Monte Carlo simulation, invariant measure

AMS subject classifications. 65C30, 60H35, 65C05

DOI. 10.1137/20M1312009

1. Introduction.

Stochastic differential equations (SDEs) are widely used in many scientific fields. Under mild assumptions, an SDE would admit a unique invariant probability measure, denoted by π. In many applications, including, but not limited, to Markov chain Monte Carlo and molecular dynamics, it is important to sample from π [1, 31]. This is usually done by either numerically integrating an SDE over very long trajectories or integrating many trajectories of the SDE over a finite time [40]. However, a numerical integrator of the SDE typically has a different invariant probability measure, denoted by $\hat{\pi}$, that depends on the time discretization [47, 48, 45]. A natural question is, How is $\hat{\pi}$ different from π? In other words, what is the quality of data sampled from a numerical trajectory of the SDE? Essentially, this is a sensitivity analysis problem of the invariant probability measure. We are interested in the robustness of π against a small change of the infinitesimal generator. This is very different from classical truncation error analysis, which is carried out for finite time intervals except in some special cases [30, 11].

Theoretically, it is well known that the distance between π and $\hat{\pi}$ can be controlled if we have good estimate of (i) the finite time truncation error and (ii) the rate of geometric ergodicity of the SDE. Estimates of this type can be made by various approaches [37, 38, 8, 5, 6]. Roughly speaking, if the truncation error over a finite time interval $[0, T]$ is $O(\epsilon)$, and the rate of geometric ergodicity is γ (i.e., speed of convergence to π is $\approx \gamma^t$ for $\gamma \in (0,1)$), then the difference between π and $\hat{\pi}$ is $O(\epsilon(1-\gamma^T)^{-1})$. (See our discussion in section 3.1

[*] Received by the editors January 10, 2020; accepted for publication (in revised form) November 10, 2020; published electronically February 1, 2021.
https://doi.org/10.1137/20M1312009
Funding: The work of the second author was partially supported by NSF DMS-1813246.

[†] Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA 01002 USA (dobson@math.umass.edu, yaoli@math.umass.edu, zhai@math.umass.edu).

for details.) However, these approaches cannot give a quantitative estimate in general, as the rate of geometric ergodicity γ estimated by rigorous approaches is usually very far from being sharp. Many approaches such as the Lyapunov function method can only rigorously show that the speed of convergence is $\approx \gamma^t$ for *some* $\gamma < 1$ [39, 19, 20]. Looking into the proof more carefully, one can easily find that this γ has to be extremely close to 1 to make the proof work. This gives a very large $(1-\gamma^T)^{-1}$ and makes rigorous estimates difficult to use in practice. To the best of our knowledge, quantitative estimates of convergence rate can only be proved for a few special cases like stochastic gradient flow and Langevin dynamics [16, 7, 2].

The aim of this paper is to provide some algorithms to numerically estimate the distance between π and $\hat{\pi}$. The finite time truncation error over a time interval $[0, T]$ is estimated by using extrapolation, which is a common practice in numerical analysis. The main novel part is the estimation of the rate of contraction of the transition kernel. Traditional approaches for computing the rate of geometric ergodicity are either computing the principal eigenvalue of the discretized generator or estimating the decay rate of correlation. The eigenvalue method works well in low dimension but faces a significant challenge if the SDE is in dimension ≥ 3. The correlation decay is difficult to estimate as well because a correlation has exponentially small expectation and large variance. One needs a huge amount of samples to estimate it effectively. In addition, exponential decay of correlation with respect to an ad hoc observable is usually not very convincing. In this paper, we propose to estimate the rate of contraction of the transition kernel by using a coupling technique.

Coupling methods have been used in rigorous proofs for decades [34, 35, 15, 42]. The idea is to run two trajectories of a random process $\boldsymbol{X} = \{X_t\}_{t\geq 0}$ where one is from a given initial distribution and the other is stationary. A suitable joint distribution, called a coupling, is constructed in the product space, such that two marginals of this joint process are the original two trajectories. If after some time, the two processes stay together with high probability, then the law of X_t must be very close to its invariant probability measure. It is well known that the coupling lemma gives bounds of both total variation norm and some 1-Wasserstein-type distances. In this paper, we use the coupling method numerically. If two numerical trajectories meet each other, they are coupled and evolve together after coupling. By the coupling lemma, the contraction rate of the transition kernel can be estimated numerically by computing the probability of successful coupling, which follows from running a Monte Carlo simulation. Together with the finite time error, we can estimate the distance between π and $\hat{\pi}$. The main advantage of a coupling method is that the estimation of coupling time distribution does not rely on spatial discretization. And many coupling strategies work well for high dimensional problems. For example, with reflection coupling, the coupling probability of two Brownian motions is independent of their dimension [35]. In this paper, we demonstrate our technique on an SDE system in $\mathbb{R}^{80}$ in section 4.5.

We provide two sets of algorithms, one for a quantitative upper bound and the other for a rough, but quick, estimate. To get the quantitative upper bound, one needs an upper bound of the contraction rate of 1-Wasserstein distance for all pairs of initial values starting from a certain compact set $\Omega \times \Omega$. This is done by applying extreme value theory. More precisely, we uniformly sample initial values from $\Omega \times \Omega$ and compute the contraction rate by using the coupling method. Then the upper bound of such a contraction rate can be obtained by numerically fitting a generalized Pareto distribution (GPD) [9, 3]. In practice, one may want a low cost estimate for the quality of samples. Hence we provide a "rough estimate" that only uses the exponential tail of the coupling probability as the rate of contraction of the generator after a given time T. This rough estimate differs from the true upper bound by an unknown constant, but it is more efficient and works well empirically.

Our coupling method can be applied to SDEs with degenerate random terms after suitable modifications. This is done by comparing the overlap of the probability density functions after two or more steps of the numerical scheme. Our approach is demonstrated on a Langevin dynamics example in section 4.3. It is known from [16, 7] that a suitable mixture of reflection coupling and synchronous coupling can be used for Langevin equation. We find that this approach can be successfully combined with the "maximal coupling" for the numerical scheme. However, for SDEs with very degenerate noise, using coupling methods remains a great challenge.

We test our algorithm with a few different examples, from simple to complicated. The sharpness of our algorithm is checked by using a "ring density example" whose invariant probability density function can be explicitly given. Then we demonstrate the use of our coupling method under degenerate noise by working with a four-dimensional (4D) Langevin equation. Next we show two examples whose numerical invariant probability $\hat{\pi}$ differs significantly from true invariant probability measure π. One is an asymmetric double well potential whose transition kernel has a slow rate of convergence. The other example is the Lorenz 96 model whose finite time truncation error is very difficult to control due to intensive chaos. Finally, we study a coupled Fitzhugh–Nagumo oscillator model proposed in [12, 27] to demonstrate that our algorithm works reasonably well in high dimensional problems.

The organization of this paper is as follows. Section 2 serves as the probability preliminary, in which we review some necessary background about the coupling method, SDEs, and numerical SDE schemes. The main algorithm is developed in section 3. All numerical examples are demonstrated in section 4. Section 5 is the conclusion.

2. Probability preliminary.

In this section, we provide some necessary probability preliminaries for this paper, which are about the coupling method, SDEs, numerical SDEs, and convergence analysis.

2.1. Coupling.

This subsection provides the definition of coupling of random variables and Markov processes.

Definition 2.1 (coupling of probability measures). *Let $\mathbb{P}$ and $\mathbb{P}'$ be two probability measures on a probability space $(\Omega, \mathcal{F})$. A probability measure γ on $(\Omega \times \Omega, \mathcal{F} \times \mathcal{F})$ is called a coupling of $\mathbb{P}$ and $\mathbb{P}'$ if two marginals of γ coincide with $\mathbb{P}$ and $\mathbb{P}'$, respectively.*

The definition of coupling can be extended to any two random variables that take values in the same state space.

Definition 2.2 (Markov coupling). *A Markov coupling of two Markov processes $\boldsymbol{X} = \{X_t\}_{t\geq 0}$ and $\boldsymbol{Y} = \{Y_t\}_{t\geq 0}$ with the same transition kernel P is a Markov process $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}) = \{(\tilde{X}_t, \tilde{Y}_t)\}_{t\geq 0}$ on the product state space $V \times V$ such that*

(i) *the marginal processes $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ are Markov processes with transition kernel P, and*
(ii) *if $\tilde{X}_s = \tilde{Y}_s$, we have $\tilde{X}_t = \tilde{Y}_t$ for all $t > s$.*

There are many ways to construct a Markov coupling between two processes. For example, let P be the transition kernel of a Markov chain $\boldsymbol{X}$ on a countable state space V, the following transition kernel Q for $(\tilde{X}_t, \tilde{Y}_t)$ on $V \times V$ such that

$$Q^t((x_1, y_1), (x_2, y_2)) = \begin{cases} P^t(x_1, x_2)P^t(y_1, y_2) & \text{if } x_1 \neq y_1, \\ P^t(x_1, x_2) & \text{if } x_1 = y_1 \text{ and } x_2 = y_2, \\ 0 & \text{if } x_1 = y_1 \text{ and } x_2 \neq y_2 \end{cases}$$

is called the *independent coupling*. Paths of the two marginal processes are independent until they first meet, after which they are identical. In the rest of this paper, unless otherwise specified, we only consider Markov couplings.

2.2. Wasserstein distance and total variation distance.

We use the following metrics in our coupling estimates.

Definition 2.3 (Wasserstein distance). *Let d be a metric on the state space V. For probability measures μ and ν on V, the Wasserstein distance (also known as Monte–Kantorovich distance) between μ and ν for d is given by*

$$\begin{aligned} d_w(\mu, \nu) &= \inf\{\mathbb{E}_\gamma[d(x, y)] \,:\, \gamma \textit{ is a coupling of } \mu \textit{ and } \nu\} \\ &= \inf\left\{\int d(x, y)\gamma(dx, dy)) \,:\, \gamma \textit{ is a coupling of } \mu \textit{ and } \nu\right\}. \end{aligned}$$

For the discussion in our paper, unless otherwise specified, we will use the Wasserstein distance with the metric

$$d(x, y) = \min\{1, \|x - y\|\}, \quad x, y \in \mathbb{R}^n. \tag{2.1}$$

Definition 2.4 (total variation distance). *Let μ and ν be probability measures on $(\Omega, \mathcal{F})$. The total variation distance of μ and ν is given by*

$$d_{TV}(\mu, \nu) = \|\mu - \nu\|_{TV} := \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|.$$

2.3. Coupling lemma.

In this subsection, we provide inequalities for the approximation of the coupling time for Markov processes.

Definition 2.5 (coupling time). *The coupling time of a Markov coupling $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}) = \{(\tilde{X}_t, \tilde{Y}_t)\}_{t \geq 0}$ is a random variable given by*

$$\tau_c = \tau_c(\boldsymbol{X}, \boldsymbol{Y}) := \inf\{t \geq 0 : \tilde{X}_t = \tilde{Y}_t\}. \tag{2.2}$$

Definition 2.6 (successful coupling). *A coupling $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}) = \{(\tilde{X}_t, \tilde{Y}_t)\}_{t \geq 0}$ of Markov processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ is said to be successful if*

$$\mathbb{P}(\tau_c(\boldsymbol{X}, \boldsymbol{Y}) < \infty) = 1$$

or equivalently,

$$\lim_{T \to \infty} \mathbb{P}(\tau_c(\boldsymbol{X}, \boldsymbol{Y}) > T) = 0.$$

For all Markov couplings, we have the following two coupling inequalities.

Lemma 2.7 (coupling inequality w.r.t. the total variation distance). *For a Markov coupling* $(\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{Y}})$ *with deterministic initial condition* $(\widetilde{X}_0, \widetilde{Y}_0) = (x, y)$, *we have*

$$\mathbb{P}(\tau_c(\boldsymbol{X}, \boldsymbol{Y}) > T) = \mathbb{P}(\widetilde{X}_T \neq \widetilde{Y}_T) \geq d_{TV}\left(P^T(x, \cdot), P^T(y, \cdot)\right).$$

Proof. For any $A \in \mathcal{F}$,

$$\begin{aligned} |P^T(x, A) - P^T(y, A)| &= |\mathbb{P}[\widetilde{X}_T \in A] - \mathbb{P}[\widetilde{Y}_T \in A]| \\ &= |\mathbb{P}[\{\widetilde{X}_T \in A\} \cap \{\widetilde{X}_T \neq \widetilde{Y}_T\}] - \mathbb{P}[\{\widetilde{Y}_T \in A\} \cap \{\widetilde{X}_T \neq \widetilde{Y}_T\}]| \\ &\leq \mathbb{P}[\widetilde{X}_T \neq \widetilde{Y}_T], \end{aligned}$$

where the second equality follows from canceling the probability

$$\mathbb{P}[\widetilde{X}_T = \widetilde{Y}_T \in A] = \mathbb{P}[\{\widetilde{X}_T \in A\} \cap \{\widetilde{X}_T = \widetilde{Y}_T\}] = \mathbb{P}[\{\widetilde{Y}_T \in A\} \cap \{\widetilde{X}_T = \widetilde{Y}_T\}].$$

By the arbitrariness of $A \in \mathcal{F}$, the lemma is proved. ■

Lemma 2.8 (coupling inequality w.r.t. the Wasserstein distance). *For a Markov coupling* $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$ *with deterministic initial condition* $(\widetilde{X}_0, \widetilde{Y}_0) = (x, y)$ *and the Wasserstein distance induced by the distance given in* (2.1), *we have*

$$\mathbb{P}(\tau_c((\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})) > T) = \mathbb{P}(\widetilde{X}_T \neq \widetilde{Y}_T) \geq d_w\left(P^T(x, \cdot), P^T(y, \cdot)\right).$$

Proof. By the definition of the Wasserstein distance,

$$\begin{aligned} d_w(P^T(x, \cdot), P^T(y, \cdot)) &\leq \int d(\xi, \eta)\mathbb{P}((\widetilde{X}_T, \widetilde{Y}_T) \in (d\xi, d\eta)) \\ &= \int_{\{\xi \neq \eta\}} d(\xi, \eta)\mathbb{P}((\widetilde{X}_T, \widetilde{Y}_T) \in (d\xi, d\eta)) \\ &\leq \int_{\{\xi \neq \eta\}} \mathbb{P}((\widetilde{X}_T, \widetilde{Y}_T) \in (d\xi, d\eta)) \\ &= \mathbb{P}(\widetilde{X}_T \neq \widetilde{Y}_T), \end{aligned}$$

where $d(x, y)$ is the specific distance given in (2.1). ■

2.4. Stochastic differential equations. We consider the following SDE for the process $\boldsymbol{X} = \{X_t\}$ with initial condition $X_0 = x_0$ that is measurable with respect to $\mathcal{F}_0 = \sigma\{B(0)\}$:

$$dX_t = f(X_t)dt + \sigma(X_t)dW_t\,, \tag{2.3}$$

where $f(X_t)$ is a continuous vector field in $\mathbb{R}^n$, $\sigma(X_t)$ is an $n \times m$ matrix-valued function, and dW_t is the white noise in $\mathbb{R}^m$. The following theorem is well known for the existence and uniqueness of the solution of (2.3) [36].

Theorem 2.9. *Assume that there are two positive constants* K_1 *and* K_2 *such that the two functions* f *and* σ *in* (2.3) *satisfy*

1. *(Lipschitz condition) for all $x, y \in \mathbb{R}^n$ and $t \in [t_0, T]$*

$$|f(x) - f(y)|^2 + |\sigma(x) - \sigma(y)|^2 \leq K_1|x - y|^2; \tag{2.4}$$

2. *(linear growth condition) for all $x, y \in \mathbb{R}^n$ and $t \in [t_0, T]$*

$$|f(x)|^2 + |\sigma(x)|^2 \leq K_2(1 + |x|^2). \tag{2.5}$$

Then there exists a unique solution $X(t)$ to (2.3) in $\mathcal{M}^2([t_0, T]; R^n) = \{g : g(t)$ is $\mathcal{F}_t$-adapted and $\mathbb{E}(\int_{t_0}^T |g(t)|^2\, dt) < \infty\}$.

In addition, we assume that $\boldsymbol{X}$ admits a unique invariant probability measure π. The existence and uniqueness of π usually follow from some drift condition plus some suitable irreducibility conditions [28, 22].

2.5. Numerical SDE. In this subsection, we talk about the numerical scheme we use for sampling SDE (2.3). The Euler–Maruyama approximation $\hat{X}_t^h$ of the solution $\boldsymbol{X}$ of (2.3) is given by

$$\hat{X}_t^h = \hat{X}_{t_{k-1}}^h + f\left(\hat{X}_{t_{k-1}}^h\right)(t - t_{k-1}) + \sigma\left(\hat{X}_{t_{k-1}}^h\right)(B(t) - B(t_{k-1})), \tag{2.6}$$

where $\hat{X}_0^h = x_0$, $t_k = t_0 + kh$, $t \in [t_{k-1}, t_k]$, and $B(t) - B(t_{k-1}) \sim \mathrm{N}(0, t - t_{k-1})$ and $B(t_j) - B(t_{j-1}) \sim \mathrm{N}(0, h), j = 1, 2, \ldots, k-1$, are mutually independent Gaussian random vectors.

We have the following convergence rate for the Euler–Maruyama approximation (see [36]).

Theorem 2.10. *Assume that (2.3) satisfies the Lipschitz condition (2.4) and the linear growth condition (2.5). Let $\boldsymbol{X}$ be the unique solution of (2.3), and let $\hat{X}_t^h$ be the Euler–Maruyama approximation for $t \in [t_0, T]$. Then*

$$\mathbb{E}\left(\sup_{t_0 \leq t \leq T} |\hat{X}_t^h - X_t|^2\right) \leq Ch, \tag{2.7}$$

where C is a constant depending only on K_1, K_2, t_0, T, and x_0.

Namely, the Euler–Maruyama approximation provides a convergence rate of order 1/2.

A commonly used improvement of the Euler–Maruyama scheme is called the Milstein scheme, which reads

$$\hat{X}_t^h = \hat{X}_{t_{k-1}}^h + f_k\left(\hat{X}_{t_{k-1}}^h\right)\Delta t + \sigma\left(\hat{X}_{t_{k-1}}^h\right)\Delta B + Z, \tag{2.8}$$

where $\Delta t = t - t_{k-1}$, $\Delta B = B(t) - B(t_{k-1})$ and Z is a vector with components

$$Z_i = \sum_{l=1}^n \sum_{j,k=1}^m \frac{\partial \sigma_{ij}\left(\hat{X}_{t_{k-1}}^h\right)}{\partial x_l} \sigma_{lk}\left(\hat{X}_{t_{k-1}}^h\right) \int_{t_k}^t B^k(s)\, dB^j(s).$$

Under suitable assumptions of Lipschitz continuity and linear growth conditions for some functions of the coefficients f and σ, the Milstein scheme [29] is an order 1 strong approximation.

Theorem 2.11 ([29]). *Under suitable assumptions, we have the following estimate for the Milstein approximation $\hat{X}_t^h$:*

$$\mathbb{E}\left(|\hat{X}_t^h - X_t|\right) \leq Kh,$$

where K is a constant independent of h.

It is easy to see that when $\sigma(X_t)$ is a constant matrix, the Euler–Maruyama scheme and the Milstein scheme coincide. In other words the Euler–Maruyama scheme for constant $\sigma(X_t)$ also has a convergence rate of order 1. There are also strong approximations of order 1.5 or 2 that are much more complicated to implement. We refer interested readers to [29].

2.6. Extreme value theory. This subsection introduces some extreme value theory that is relevant to materials in this paper.

Definition 2.12 (generalized Pareto distribution). *A random variable Y is said to follow a GPD if its cumulative distribution function is given by*

$$F_{\xi,\zeta}(x) = \begin{cases} 1 - \left(1 + \frac{\xi x}{\zeta}\right)^{-1/\xi} & \textit{if } \xi \neq 0, \\ 1 - \exp\left(-\frac{x}{\zeta}\right) & \textit{if } \xi = 0, \end{cases}$$

where $\zeta > 0$, and $x \geq 0$ if $\xi \geq 0$ and $0 \leq x \leq -\zeta/\xi$ if $\xi < 0$.

The GPD is used to model the so-called peaks over threshold distribution, that is, the part of a random variable over a chosen threshold u, or the tail of a distribution. Specifically, for a random variable X with cumulative distribution function $F(x)$, consider the random variable $X - u$ conditioned on the threshold u being exceeded. Its conditional distribution function is called the *conditional excess distribution function* and is denoted by

$$F_u(x) = P(X - u \leq x | X > u) = \frac{P(\{X - u \leq x\} \cap \{X > u\})}{P(X > u)} = \frac{F(x+u) - F(u)}{1 - F(u)}.$$

Extreme value theory can be used to prove the following theorem.

Theorem 2.13 (see [4, 44]). *For a large class of distributions (e.g., uniform, normal, log-normal, t, F, gamma, beta distributions), there is a function $\zeta(u)$ such that*

$$\lim_{u \to \bar{x}} \sup_{0 \leq x < \bar{x} - u} |F_u(x) - F_{\xi,\zeta(u)}(x)| = 0,$$

where $\bar{x} = \sup\{x | F(x) < 1\}$ is the rightmost point of the distribution.

This theorem shows that the conditional distribution of peaks over threshold can be approximated by the GPD. In this paper, we will employ this method to estimate the upper bound of quantities of interest. In our algorithm, we use the maximum likelihood estimation method to fit the parameters ζ and ξ and then use them to compute the estimates. To be specific, for any $u > 0$, the following formula is applied:

$$F_{\xi,\zeta}(x) \approx \frac{F(x+u) - F(u)}{1 - F(u)} \approx \frac{N_{x+u}/N - N_u/N}{1 - N_u/N} = \frac{N_{x+u} - N_u}{N - N_u},$$

where N_u is the number of samples that is less than or equal to u and N is the sample size, namely, here we use the empirical probabilities to approximate the cumulative probability function values. In practice, this can be done using the MATLAB function *gpfit*. Note that parameters ζ and ξ are dependent on the threshold u, so one need to manually choose a suitable u. The criterion of choosing a suitable threshold is that (i) u should be large enough such that only less than 1000 samples are greater than u, and (ii) the output of *gpfit*, i.e., ζ and ξ, are stabilized with increasing u until there are too few available samples above u.

3. Description of algorithm.

3.1. Decomposition of error terms. Let $\boldsymbol{X} = \{X_t\}$ and $\boldsymbol{X}^h = \{\hat{X}^h_t\}$ be the stochastic processes given by (2.3) and a numerical approximation with step size h, respectively. Let P and $\hat{P}$ be the two corresponding transition kernels. Hence in $\hat{X}^h_t$, t only takes values $0, h, 2h, \cdots$. Let $T > 0$ be a fixed constant. Let d_w be the 1-Wasserstein distance of probability measures induced by distance

$$d(x,y) = \min\{1, \|x-y\|\}, \quad x, y \in \mathbb{R}^n,$$

unless otherwise specified.

Denote the invariant probability measures of $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}^h$ by π and $\hat{\pi}$, respectively. The quantity that we are interested in is $\mathrm{d}_w(\pi, \hat{\pi})$. We need to know whether $\hat{\pi}$ is a good approximation of π for a reasonable time step size h. In this paper, we set the threshold at 0.05. If the computation shows that $\mathrm{d}_w(\pi, \hat{\pi}) < 0.05$ for a numerical scheme with a reasonable time step size h, this numerical scheme is considered to be trustable.

The following decomposition follows easily by the triangle inequality and the invariance. (This is motivated by [25]. Similar approaches are also reported in [46, 41].)

$$\mathrm{d}_w(\pi, \hat{\pi}) \le \mathrm{d}_w\left(\pi P^T, \pi \hat{P}^T\right) + \mathrm{d}_w\left(\pi \hat{P}^T, \hat{\pi} \hat{P}^T\right) . \tag{3.1}$$

If the transition kernel $\hat{P}^T$ has enough contraction such that

$$\mathrm{d}_w\left(\pi \hat{P}^T, \hat{\pi} \hat{P}^T\right) \le \alpha \mathrm{d}_w(\pi, \hat{\pi}) \,,$$

for some $\alpha < 1$, we have

$$\mathrm{d}_w(\pi, \hat{\pi}) \le \frac{\mathrm{d}_w\left(\pi P^T, \pi \hat{P}^T\right)}{1-\alpha} \,.$$

In other words the distance $\mathrm{d}_w(\pi, \hat{\pi})$ can be estimated by computing the finite time error and the speed of contraction of $\hat{P}^T$. Theoretically, the finite time error can be given by the strong approximation of the truncation error of the numerical scheme of (2.3). The second term comes from the geometric ergodicity of the Markov process $\hat{\boldsymbol{X}}^h$. As discussed in the introduction, except in some special cases, the rate of geometric ergodicity of $\hat{\boldsymbol{X}}^h$ cannot be estimated sharply. As a result one can only have an α that is extremely close to 1. Therefore, we need to look for suitable numerical estimators of the two terms in (3.1).

The other difficulty comes from the fact that both $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}^h$ are defined on an unbounded domain. However, for a very large class of SDEs, large deviations theory guarantees that the mass of both π and $\hat{\pi}$ should concentrate near the global attractor of the deterministic part of (2.3) (i.e., the ODE $X_t' = f(X_t)$) [28, 17]. Similar concentration estimates can be made by many different approaches [33, 23]. Therefore, we assume that there exists a compact set Ω and a constant $0 < \epsilon \ll 1$ such that

$$\pi(\Omega^c) < \epsilon\,, \quad \hat{\pi}(\Omega^c) < \epsilon\,, \quad \pi\hat{P}^T(\Omega^c) < \epsilon\,. \tag{3.2}$$

In practice, Ω can be chosen to be a set that contains all samples of a very long trajectory of $\hat{\boldsymbol{X}}^h$, and ϵ is the reciprocal of the length of this trajectory. This ϵ is usually significantly smaller than all other error terms. Algorithm 1 needs to run a long trajectory. Hence set Ω (that will be needed in Algorithm 2) can be obtained after running Algorithm 1.

This allows us to estimate the contraction rate α for initial values in a compact set. Let Γ be the optimal coupling plan such that

$$\mathrm{d}_w(\pi, \hat{\pi}) = \int_{\mathbb{R}^n\times\mathbb{R}^n} d(x,y)\Gamma(\mathrm{d}x, \mathrm{d}y)\,.$$

Consider a Markov coupling of two trajectories of $\hat{\boldsymbol{X}}^h$, and let $\hat{P}_c$ denote the corresponding transition kernel on $\mathbb{R}^n \times \mathbb{R}^n$. By the assumption of Ω, we have

$$\begin{aligned}\mathrm{d}_w\left(\pi\hat{P}^T, \hat{\pi}\hat{P}^T\right) &\le \int_{\mathbb{R}^n\times\mathbb{R}^n} d(x,y)\Gamma(\hat{P}_c)^T(\mathrm{d}x,\mathrm{d}y)\\ &\le 2\epsilon + \int_{\Omega\times\Omega} d(x,y)\Gamma(\hat{P}_c)^T(\mathrm{d}x,\mathrm{d}y)\\ &\le 2\epsilon + \alpha_\Omega \int_{\Omega\times\Omega} \mathrm{d}(x,y)\Gamma(\mathrm{d}x,\mathrm{d}y)\\ &\le 2\epsilon + \alpha_\Omega \mathrm{d}_w(\pi,\hat{\pi})\,,\end{aligned} \tag{3.3}$$

where α_Ω is the minimum contracting rate of $(\hat{P}_c)^T$ on $\Omega\times\Omega$ such that

$$\alpha_\Omega = \sup_{(x,y)\in\Omega\times\Omega} \frac{\mathrm{d}_w\left(\delta_x\hat{P}^T, \delta_y\hat{P}^T\right)}{d(x,y)}\,,$$

where δ_x and δ_y are the two one-point distributions concentrated at x and y, respectively, and $\delta_x\hat{P}^T$ and $\delta_y\hat{P}^T$ are two marginal distributions of $\delta_{(x,y)}(\hat{P}_c)^T$.

Combine (3.1) and (3.3), we have

$$\mathrm{d}_w\left(\pi, \hat{\pi}\right) \le \frac{\mathrm{d}_w(\pi P^T, \pi\hat{P}^T) + 2\epsilon}{1-\alpha_\Omega}\,. \tag{3.4}$$

It remains to discuss the choice of T. We find that the result does not depend sensitively on T as long as T is in a suitable range, such that $(1-\alpha_\Omega)^{-1} = O(1)$. In practice, we choose T such that $(1-\alpha_\Omega)^{-1}$ is between 1.5 and 3. Smaller (resp., larger) T works better if $\mathrm{d}_w(\pi P^T, \pi\hat{P}^T)$ grows faster (resp., slower) with increasing T.

3.2. Estimator of error terms.

From (3.4), we need to numerically estimate the finite time error $\mathrm{d}_w(\pi P^T, \pi\hat{P}^T)$ and the contraction rate α_Ω. We propose the following approach to estimate these two quantities.

Extrapolation for finite time error. There are some known analytical results about the stability error $\mathrm{d}_w(\pi P^T, \pi\hat{P}^T)$ in 1-Wasserstein norm (or other weak norms), which usually give the order of the finite time error in terms of h with an unknown prefactor [14, 10, 18, 43]. But these estimates are not quantitative because the prefactor is usually unknown. In this paper, we use an extrapolation to numerically estimate this error term. By the definition of 1-Wasserstein distance, we have

$$\mathrm{d}_w(\pi P^T, \pi\hat{P}^T) \le \int_{\mathbb{R}^n\times\mathbb{R}^n} d(x,y)\gamma(\mathrm{d}x, \mathrm{d}y)$$

for any coupling measure γ. A suitable choice of γ that can be sampled easily is $\pi^2(P^T \circ \hat{P}^T)$, where π^2 is the coupling measure of π on the "diagonal" of $\mathbb{R}^n \times \mathbb{R}^n$ that is supported by the hyperplane

$$\{(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{2n} \,|\, \mathbf{y} = \mathbf{x}\}\,,$$

such that $\pi^2(\{(\mathbf{x},\mathbf{x}) \,|\, \mathbf{x} \in A\}) = \pi(A)$.

Theoretically, we can sample an initial value from π, run $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}^h$ up to time T, and calculate $d(X_T, \hat{X}^h_T)$. However, we do not have exact expressions for π and X_t. Hence in the estimator, we use $\hat{\pi}$ to replace π and use extrapolation to estimate $d(X_T, \hat{X}^h_T)$. We need to assume that $\|\hat{\pi} - \pi\|_{TV}$ is a small quantity such that $\|\hat{\pi} - \pi\|^2_{TV} \ll \|\hat{\pi} - \pi\|_{TV}$. Then

$$\int_{\mathbb{R}^n\times\mathbb{R}^n} d(x,y)\pi^2(P^T \circ \hat{P}^T)(\mathrm{d}x, \mathrm{d}y)$$

can be approximated by

$$\int_{\mathbb{R}^n\times\mathbb{R}^n} d(x,y)\hat{\pi}^2(P^T \circ \hat{P}^T)(\mathrm{d}x, \mathrm{d}y)\,.$$

This approximation causes a very small error,

$$\begin{aligned}\|\pi - \hat{\pi}\|_{TV} \times \Bigg(&\int_{\mathbb{R}^n\times\mathbb{R}^n} d(x,y)(\nu^+)^2(P^T \circ \hat{P}^T)(\mathrm{d}x, \mathrm{d}y) \\ &- \int_{\mathbb{R}^n\times\mathbb{R}^n} d(x,y)(\nu^-)^2(P^T \circ \hat{P}^T)(\mathrm{d}x, \mathrm{d}y)\Bigg)\,,\end{aligned} \tag{3.5}$$

where $\nu = (\|\pi - \hat{\pi}\|_{TV})^{-1}(\pi - \hat{\pi})$ is a renormalized signed measure with zero mass and total variation 1. The two terms are both at the same magnitude as $d_w(\pi, \hat{\pi})$, hence (3.5) gives a higher order error. We assume that this error term is negligible.

The integral

$$\int_{\mathbb{R}^n\times\mathbb{R}^n} d(x,y)\hat{\pi}^2(P^T \circ \hat{P}^T)(\mathrm{d}x, \mathrm{d}y)$$

is computable. It can be estimated by sampling $\mathrm{d}(X_T, \hat{X}^h_T)$ such that $X_0 = \hat{X}_0 \sim \hat{\pi}$. The distance $d(X_T, \hat{X}^h_T)$ can be obtained by extrapolating $d(\hat{X}^h_T, \hat{X}^{2h}_T)$, where $\hat{X}^{2h}_T$ is the random process of the same numerical scheme with the same noise term but $2h$ time step size. For the same noise, we mean the sum of noise terms of two h-steps should be equal to the noise term of one $2h$-step. Take the Euler-Maruyama scheme as an example. The update of $\hat{X}_t^{2h}$ and $\hat{X}_t^h$ should follow

$$\hat{X}_{t+h}^h = \hat{X}_t^h + f(\hat{X}_t^h)h + \sigma(\hat{X}_t^h)\sqrt{h}N_1,$$
$$\hat{X}_{t+2h}^h = \hat{X}_{t+h}^h + f(\hat{X}_{t+h}^h)h + \sigma(\hat{X}_{t+h}^h)\sqrt{h}N_2,$$

and

$$\hat{X}_{t+2h}^{2h} = \hat{X}_t^h + f(\hat{X}_t^h)h + \sigma(\hat{X}_t^h)\sqrt{2h}(N_1 + N_2)/\sqrt{2},$$

respectively, where N_1 and N_2 are two independent standard normal random variables. See Algorithm 1 for more details of the implementation.

When N is sufficiently large, $\mathbf{x}_1, \dots, \mathbf{x}_N$ in Algorithm 1 are from a long trajectory of the time-T skeleton of $\hat{X}_T^h$. Hence $\mathbf{x}_1, \dots, \mathbf{x}_N$ are approximately sampled from $\hat{\pi}$. The error term $y_i = cd(\hat{X}_T^h, \hat{X}_T^{2h})$ for $\hat{X}_0^h = \hat{X}_0^{2h} = \mathbf{x}_i$ estimates $d(X_T, \hat{X}_T^h)$. Therefore,

$$\frac{1}{N}\sum_{i=1}^N y_i$$

estimates the integral

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} d(x,y)\hat{\pi}^2\left(P^T \circ \hat{P}^T\right)(\mathrm{d}x, \mathrm{d}y)\,,$$

which is an upper bound of $\mathrm{d}_w(\pi P^T, \pi\hat{P}^T)$.

The constant c in Algorithm 1 is chosen based on the strong order of accuracy. If the numerical scheme is a strong approximation with order p, the error $\mathrm{d}(X_T, \hat{X}_T^h)$ is $O(h^p)$. Then we have

$$\mathrm{d}(X_T, \hat{X}_T^{2h}) \approx 2^p \mathrm{d}(X_T, \hat{X}_T^h).$$

This gives

$$(2^p - 1)\mathrm{d}(X_T, \hat{X}_T^h) \approx \mathrm{d}(X_T, \hat{X}_T^{2h}) - \mathrm{d}(X_T, \hat{X}_T^h) \le \mathrm{d}(\hat{X}_T^h, \hat{X}_T^{2h})\,,$$

which gives $c = (2^p - 1)^{-1}$.

Algorithm 1 Estimate finite time error

```
**Input:** Initial value $\mathbf{x}_0$
**Output:** An estimator of $\mathrm{d}_w(\pi P^T, \pi\hat{P}^T)$
Run the numerical trajectory for some time $T_0$ to "burn in." Let $\mathbf{x}_1 = \hat{X}_{T_0}^h$.
**for** i = 1 to N **do**
    Using the same noise, simulate $\hat{\boldsymbol{X}}^h$ and $\hat{\boldsymbol{X}}^{2h}$ with initial value $\mathbf{x}_i$ up to $t = T$.
    Let $y_i = cd(\hat{X}_T^h, \hat{X}_T^{2h})$, where $c = (2^p - 1)^{-1}$, $p$ is the strong order of convergence of the numerical scheme.
    Let $\mathbf{x}_{i+1} = \hat{X}_T^h$.
**end for**
Return $\frac{1}{N}\sum_{i=1}^N y_i$
```

One advantage of Algorithm 1 is that it can run together with the Monte Carlo sampler. The trajectory of $\hat{X}^{2h}_t$ cannot be recycled. But the trajectory of $\hat{X}^h_t$ can be used to estimate either the invariant density or the expectation of an observable.

Coupling for contraction rate. The idea of estimating α_Ω is to use coupling. We can construct a Markov process $\hat{Z}_t = (\hat{X}^{(1)}_t, \hat{X}^{(2)}_t)$ such that $\hat{Z}_t$ is a Markov coupling of $\hat{X}^{(1)}_t$ and $\hat{X}^{(2)}_t$. Then as introduced in section 2, the first passage time to the "diagonal" hyperplane $\{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{2n} \,|\, \mathbf{y} = \mathbf{x}\}$ is the *coupling time*, which is denoted by τ_c. It then follows from Lemma 2.8 that

$$\mathrm{d}_w(\delta_x \hat{P}^T, \delta_y \hat{P}^T) \le \mathbb{P}[\tau_c > T] \,.$$

Then we can use extreme value theory to estimate α_Ω. The idea is to uniformly sample initial values (x, y) from $\Omega \times \Omega$, and define $\beta(x, y) := \mathbb{P}[\tau_c > T]/d(x, y)$. Then β is actually a random variable whose sample can be easily computed. Assume M samples starting from each (x, y) are simulated, and K of them have coupled before time T. An estimator of $\beta(x, y)$ is $K/(d(x, y)M)$. We use extreme value theory to estimate an upper bound for β and denote it by α_Ω. See Algorithm 2 for details. The threshold V in Algorithm 2 is usually chosen such that approximately 5% samples are greater than this threshold. The goal is to make (i) the empirical cumulative distribution function matches that of the resultant GPD, and (ii) the result α_Ω does not sensitively depend on the choice of the threshold.

If running successfully, Algorithms 1 and 2 give us an upper bound of $\mathrm{d}_w(\pi, \hat{\pi})$ according to (3.4), which can be used to check the quality of samples.

Construction of $\hat{Z}_t$. It remains to construct a coupling scheme that is suitable for the numerical trajectory $\hat{X}_t$. In this paper we use the following two types of couplings.

Denote two margins of $\hat{Z}_t$ by $\hat{X}^{(1)}_t$ and $\hat{X}^{(2)}_t$, respectively. A *reflection coupling* of the SDE (2.3) means the noise terms of $\hat{X}^{(1)}_t$ and $\hat{X}^{(2)}_t$ in an update are symmetric about the normal plane bisecting the line segment between their positions. For the case of a constant coefficient matrix $\sigma(X_t) = \sigma$ and using the Euler–Maruyama scheme, reflection coupling gives the following update from t to $t + h$:

$$\begin{aligned} \hat{X}^{(1)}_{t+h} &= \hat{X}^{(1)}_t + f\left(\hat{X}^{(1)}_t\right) h + \sigma\sqrt{h} N_t, \\ \hat{X}^{(2)}_{t+h} &= \hat{X}^{(2)}_t + f\left(\hat{X}^{(2)}_t\right) h + \sigma\sqrt{h}(I - 2e_t e_t^T) N_t \,, \end{aligned} \tag{3.6}$$

where N_t is a normal random variable with mean 0 and variance $h = dt$, and

$$e_t = \frac{1}{\left\|\sigma^{-1}\left(\hat{X}^{(1)}_t - \hat{X}^{(2)}_t\right)\right\|} \sigma^{-1}\left(\hat{X}^{(1)}_t - \hat{X}^{(2)}_t\right)$$

is a unit vector. It is known that reflection coupling is the optimal coupling for Brownian motion in $\mathbb{R}^n$ [35, 21]. Empirically it gives fast coupling rates for many SDEs with nondegenerate noise.

The *maximal coupling* looks for the maximal coupling probability for the next step (or next several steps) of the numerical scheme. Assume $\hat{X}^{(1)}_t$ and $\hat{X}^{(2)}_t$ are both known. Then it is easy to explicitly calculate the probability density function of $\hat{X}^{(1)}_{t+h}$ and $\hat{X}^{(2)}_{t+h}$, denoted

Algorithm 2 Estimate contraction rate

```
Input: A compact set $\Omega$
Output: An estimator of the minimal contraction rate $\alpha_\Omega$.
for i = 1 to N do
    Sample pairs $(x_i, y_i)$ uniformly from $\Omega \times \Omega$.
    Set $K_i = 0$, $r_i = 0$.
    for j = 1 to M do
        Run $\hat{Z}_t$ with initial value $(x_i, y_i)$ until $\min\{\tau_c, T\}$.
        if $\tau_c \leq T$ then
            $K_i \leftarrow K_i + 1$
        end if
    end for
    $\beta_i \leftarrow K_i/(d(x_i, y_i)M)$
end for
if $\max\{\beta_i\} \geq 1$ then
    The estimator fails. Choose better coupling algorithm or larger $T$.
else
    Let $v_i = 1/(1 - \beta_i)$ for all $1 \leq i \leq N$.
    Choose a threshold $V$.
    Use GPD to fit $\{v_i - V \,|\, v_i \geq V\}$ to get two parameters $(\zeta, \xi)$.
    if $\xi \geq 0$ then
        The estimator fails. Choose better coupling algorithm or larger $T$.
    else
        Let $v_{max} = V - \zeta/\xi$.
    end if
end if
Return $1 - 1/v_{max}$ as the estimator of $\alpha_\Omega$.
```

by $p^{(1)}(x)$ and $p^{(2)}(x)$, respectively. The update of $\hat{X}^{(1)}_{t+h}$ and $\hat{X}^{(2)}_{t+h}$ is described in Algorithm 3, which is adopted from [24, 26]. This update maximizes the probability of coupling at the next step. We adopt the name "maximal coupling" from [24].

In practice, we use reflection coupling when $\hat{X}^{(1)}_t$ and $\hat{X}^{(2)}_t$ are far away from each other, and maximal coupling when they are sufficiently close. This method is suited for discrete-time numerical schemes, as using reflection coupling alone will easily allow the two processes to miss each other. In our simulation code, the threshold of the changing coupling method is $2\sqrt{h}\|\sigma\|$. When the distance between $\hat{X}^{(1)}_t$ and $\hat{X}^{(2)}_t$ is smaller than this, we use maximal coupling. In practice, the coupling speed does not depend on this threshold sensitively as long as it is $O(\sqrt{h}\sigma)$.

3.3. A fast estimator. In practice, Algorithm 1 can be done together with the Monte Carlo sampler to compute either an observable or the invariant probability density function. The extra cost comes from simulating trajectories of $\hat{\boldsymbol{X}}^{2h}$, which takes 50% of the time of running the trajectories of $\hat{\boldsymbol{X}}^{h}$. The main overhead of the above mentioned methods is

Algorithm 3 Maximal coupling

```
Input: $\hat{X}_t^{(1)}$ and $\hat{X}_t^{(2)}$
Output: $\hat{X}_{t+h}^{(1)}$, $\hat{X}_{t+h}^{(2)}$, and $\tau_c$ if the coupling is successful.
Compute probability density functions $p^{(1)}(x)$ and $p^{(2)}(x)$.
Sample $\hat{X}_{t+h}^{(1)}$ and calculate $W = Up^{(1)}(\hat{X}_{t+h}^{(1)})$, where $U$ is a uniform random variable on $(0,1)$.
if $W \le p^{(2)}(\hat{X}_{t+h}^{(1)})$ then
    $\hat{X}_{t+h}^{(2)} = \hat{X}_{t+h}^{(1)}$, $\tau_c = t+h$, the coupling is successful.
else
    Sample $\hat{X}_{t+h}^{(2)}$ and calculate $W' = Vp^{(2)}(\hat{X}_{t+h}^{(2)})$, where $V$ is a uniform random variable on $(0,1)$.
    while $W' \le p^{(1)}(\hat{X}_{t+h}^{(2)})$ do
        Resample $\hat{X}_{t+h}^{(2)}$ and $V$. Calculate $W' = Vp^{(2)}(\hat{X}_{t+h}^{(2)})$ with new samples.
    end while
    Let $t = t+h$. The coupling is unsuccessful and $\tau_c$ is undetermined.
end if
```

Algorithm 2. Because we want a quantitative upper bound of $\mathrm{d}_w(\pi, \hat{\pi})$, the contraction rate of $\hat{P}^T$ in 1-Wasserstein space with respect to all pairs of initial points in $\Omega \times \Omega$ must be estimated. In practice, this takes a long time because one needs to run many $(100 - 1000)$ independent trajectories from each initial point to estimate the coupling probability.

In practice, if one only needs a rough estimate about the sample quality instead of a definite upper bound of the 1-Wasserstein distance, Algorithm 2 can be done in a much easier way by estimating the exponential *rate* of convergence of $\hat{P}^t$. It is usually safe to assume that the rate of exponential contraction is same for all "reasonable" initial distributions. In addition, the contraction rate is bounded from below by the exponential tail of the coupling time distribution $\mathbb{P}[\tau_c > t]$. Therefore, we only need to sample some initial points uniformly distributed in $\Omega \times \Omega$ and estimate the exponential tail of the coupling time distribution. This gives an estimate

$$\lim_{t\to\infty} \frac{1}{t} \log(\mathbb{P}_u[\tau_c > t]) = -\gamma \,,$$

where u denotes the uniform distribution on $\Omega \times \Omega$, and γ can be obtained by a linear fit of $\mathbb{P}_u[\tau_c > t]$ versus t in a log-linear plot. In other words, we have

$$\mathrm{d}_w(\pi\hat{P}^t, \hat{\pi}\hat{P}^t) \le C_{\pi,\hat{\pi}} e^{-\gamma t} \mathrm{d}_w(\pi, \hat{\pi}) \,.$$

The unknown prefactor $C_{\pi,\hat{\pi}}$ depends on π and $\hat{\pi}$. Assuming $C_{\pi,\hat{\pi}} = 1$, we have

$$\mathrm{d}_w(\pi\hat{P}^T, \hat{\pi}\hat{P}^T) \le e^{-\gamma T} \mathrm{d}_w(\pi, \hat{\pi}) \,.$$

Combining with (3.1), a rough estimate of $\mathrm{d}_w(\pi, \hat{\pi})$ is given by

$$\mathrm{d}_w(\pi, \hat{\pi}) \approx \frac{\mathrm{d}_w(\pi P^T, \pi\hat{P}^T)}{1 - e^{-\gamma T}} \,, \tag{3.7}$$

where $\mathrm{d}_w(\pi P^T, \pi\hat{P}^T)$ is estimated by Algorithm 1.

Equation (3.7) usually differs from the output of Algorithm 1, Algorithm 2, and (3.4) by an unknown multiplicative constant. However, in practice, this is usually sufficient for us to predict the quality of the Monte Carlo sampler with relatively low computational cost. In numerical examples we will show that it is usually sufficient to estimate the exponential tail of $\mathbb{P}_u[\tau_c > t]$ by running $10^4 - 10^5$ trajectories.

4. Numerical examples.

4.1. Ring density. The first example is the "ring density" example that has a known invariant probability measure. Consider the following SDE:

$$\begin{cases} dX_t = \big(-4X_t(X_t^2 + Y_t^2 - 1) + Y_t \big)\, dt + \sigma\, dW_t^{(1)}, \\ dY_t = \big(-4Y_t(X_t^2 + Y_t^2 - 1) - X_t \big)\, dt + \sigma\, dW_t^{(2)}, \end{cases} \tag{4.1}$$

where $W_t^{(1)}$ and $W_t^{(2)}$ are independent Wiener processes, and σ is the strength of the noise. The drift part of (4.1) is a gradient flow of the potential function $V(x, y) = (x^2 + y^2 - 1)^2$ plus a rotation term orthogonal to the equipotential lines of V. This rotation term does not change the invariant probability density function, which can be verified by plugging $V(x, y)$ into the Fokker–Planck equation. Hence the invariant probability measure of (4.1) has a probability density function

$$u(x, y) = \frac{1}{K} e^{-2V(x,y)/\sigma^2},$$

where $K = \pi \int_{-1}^{\infty} e^{-2t^2/\sigma^2}\, dt$ is a normalizer. We will compare the invariant probability measure of the Euler–Maruyama scheme and that of (4.1).

In our simulation, we choose $\sigma = 0.5$ and $T = 10$. The first simulation runs eight independent long trajectories up to time 1.25×10^6. Hence Algorithm 1 compares the distance between $\hat{X}_T^h$ and $\hat{X}_T^{2h}$ for 10^7 samples. Constant c equals 1 here because the Euler–Maruyama scheme is a strong approximation with accuracy $O(h)$ when σ is a constant. Algorithm 1 gives an upper bound

$$\mathrm{d}_w(\pi P^T, \pi \hat{P}^T) \le 0.00141635\,.$$

In addition, all eight trajectories are contained in the box $[-2, 2]^2$. Hence we choose $\Omega = [-2, 2]^2$ and $\epsilon = 10^{-7}$.

Then we run Algorithm 2 to get coupling probabilities up to $T = 10$. The number of initial values (x_i, y_i) is 20000. Then we run 1000 pairs of trajectories from each initial point to estimate the coupling probability. The probability that coupling has not happened before T is then divided by $d(x_i, y_i)$, which estimates an upper bound of the contraction rate

$$r_i = \frac{\mathbb{P}_{x_i,y_i}[\tau_c > T]}{d(x_i, y_i)} \ge \frac{\mathrm{d}_w(\delta_{x_i}\hat{P}^T, \delta_{y_i}\hat{P}^T)}{d(x_i, y_i)}\,.$$

Then we use GPD to fit $\{1/(1 - r_i)\}_{i=1}^{20000}$. The threshold V is chosen to be 1.48. The fitting algorithm gives parameters $\xi = -0.0419$ and $\zeta = 0.0254$. This gives $\alpha_\Omega = 0.5207$. A comparison of cumulative distribution functions of empirical data and that of the GPD fitting is demonstrated in Figure 1, top left.

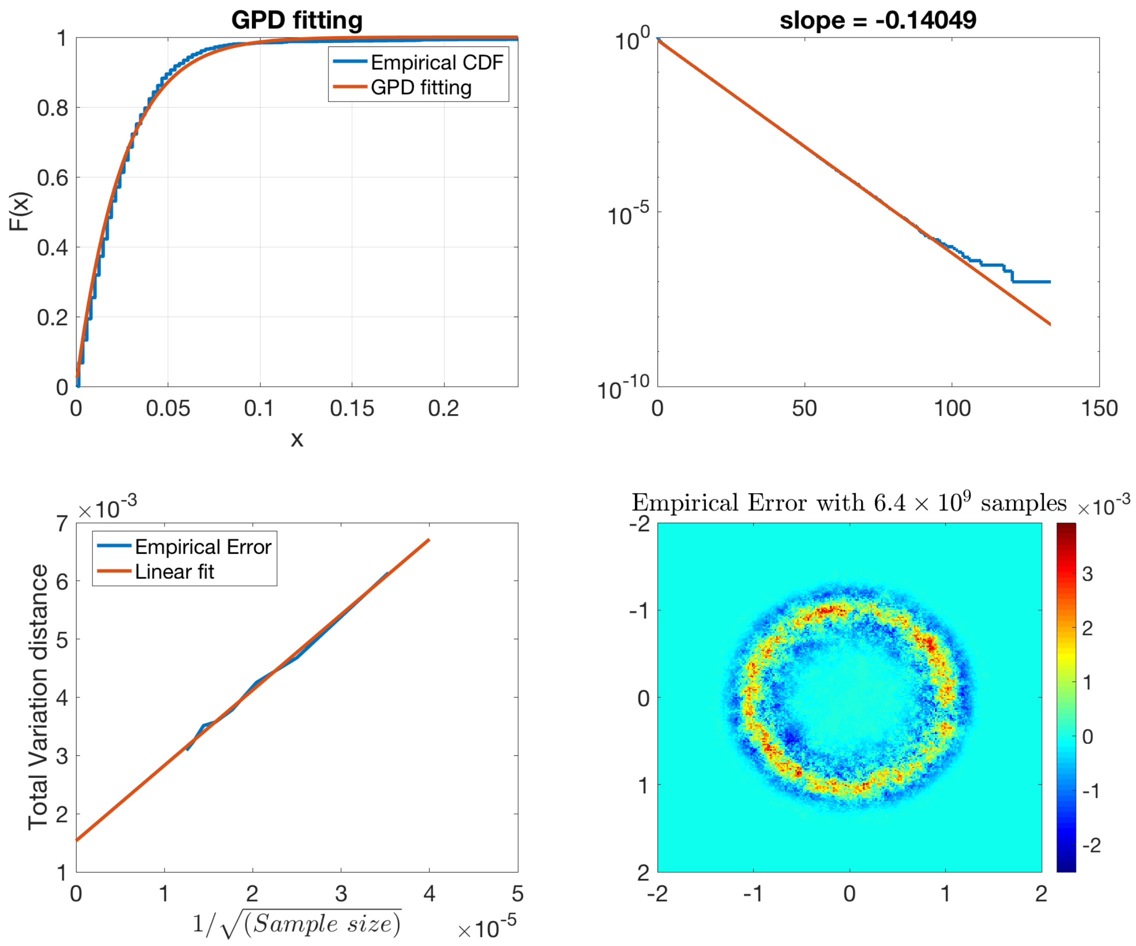

Figure 1. *Top left: A comparison of cumulative distribution functions of empirical data $\{v_i \mid v_i > V\}$ and that of the GPD. Top right: Exponential tail of $\mathbb{P}[\tau_c > t]$ versus t when initial values are uniformly sampled in $\Omega \times \Omega$. Bottom left: Linear extrapolation for the total variance distance $\|\pi - \hat{\pi}\|_{TV}$ at the infinite sample limit. Bottom right: Difference between $\hat{\pi}$ and π with 6.4×10^9 samples.*

Combining all estimates above, we obtain a bound

$$\mathrm{d}_w(\pi, \hat{\pi}) \leq 0.002955\,. \tag{4.2}$$

Since the invariant probability measure of (4.1) is known, we can check the sharpness of the bound given in (4.2). The approach we take is Monte Carlo simulation with extrapolation to infinite sample size. On a 256 × 256 grid, we use eight long trajectories to estimate the invariant probability density function of (4.1). The sample sizes of these trajectories are $8\times10^8, 1.6\times10^9, \ldots, 6.4\times10^9$. As seen in Figure 1, bottom right, the error between probability density functions of π and $\hat{\pi}$ is at the magnitude of 10^{-3}. Then we compute the total variation distance between

$$u(x,y) = \frac{1}{K} e^{-2V(x,y)/\sigma^2}$$

and the empirical probability density function at those grid points. As seen in Figure 1, bottom right, the error is inversely proportional to the square root of the sample size. Linear extrapolation shows that the total variation distance at the infinite sample limit is ≈ 0.001534. The linear extrapolation is demonstrated in Figure 1, bottom left. Since d_w is smaller than the total variation distance, the 1-Wasserstein distance $\mathrm{d}_w(\pi, \hat{\pi})$ should be no greater than 0.001534. (The total variation distance can be seen as a 1-Wasserstein distance with underlying distance $\tilde{d}(x,y) = \mathbf{1}_{\{x=y\}}$, which is larger than the distance $d(x,y)$ used throughout this paper.) Therefore, our estimation given in (4.2) is larger than the true distance between π and $\hat{\pi}$ but is reasonably sharp.

It remains to comment on the fast estimator mentioned in section 3.3. In Figure 1, top right, we draw the exponential tail of $\mathbb{P}[\tau_c > t]$ and its linear fit. The slope of the exponential tail is $\gamma = -0.14049$. When $T = 10$, we have $e^{-\gamma T} = 0.2454$. Equation (3.7) then gives an estimate

$$\mathrm{d}_w(\pi, \hat{\pi}) \approx 0.001877\,,$$

which is actually closer to the total variation distance that we have measured numerically.

4.2. Double well potential.

The second example we study is a gradient flow with respect to an asymmetric double well potential. Let

$$V(x) = \begin{cases} 6x^2 - 60 & \text{if } x \geq 4, \\ \frac{1}{4}x^4 - 2x^2 + 4 & \text{if } 0 \leq x < 4, \\ \frac{1}{4}r^4x^4 - 2r^2x^2 + 4 & \text{if } -4/r \leq x < 0, \\ 6r^2x^2 - 60 & \text{if } x < -4/r. \end{cases}$$

If $r \neq 1$, V is an asymmetric double well potential function. Note that we make $V(x)$ a quadratic function when $x \geq 4$ or $x < -4/r$, because the original quartic function has very large derivatives when $|x|$ is large, which has some undesired numerical artifacts.

Now consider the gradient flow of $V(x)$ with additive random perturbation

$$dX_t = -V'(X_t)\,dt + \sigma \mathrm{d}W_t\,.$$

It is easy to see that X_t admits an invariant probability measure π with probability density function

$$u(x) = \frac{1}{K}e^{2V(x)/\sigma^2}\,,$$

where K is a normalizer.

Because of the double well potential, trajectories from two local minima need a long time to meet with one another. Hence the speed of convergence of the law of X_t to π is slow. Much longer times are needed so that trajectories can couple in Algorithm 2. In addition, we changed the underlying distance from $|x-y|$ to $|x-y|^{0.45}$, because if two initial values are very close to each other, with some small probability one trajectory can run into a different local minimum and takes a very long time to return. As a result, for reasonably large T, if the underlying distance $|x-y|$ is used, $\hat{P}^T$ does not contract in 1-Wasserstein metric space when two initial points are very close to each other.

Model parameters are chosen to be $r = 5$, $\sigma = 1.2$, and $T = 50$. Then we choose $\Omega = [-2, 4]$, as $u(x)$ is extremely small when $x < -2$ or $x > 4$. The numerical trajectory $\hat{X}_t^h$

is obtained by running the Euler–Maruyama scheme with $h = 0.0025$. We first run Algorithm 1 with eight independent long trajectories to compare the distance between $\hat{X}^h_T$ and $\hat{X}^{2h}_T$. The length of each trajectory is 5×10^6. The constant c is still equal to 1 because the accuracy of the Euler–Maruyama scheme is $O(h)$ when σ is a constant. Algorithm 1 gives an upper bound

$$\mathrm{d}_w(\pi P^T, \pi \hat{P}^T) \leq 0.167345\,.$$

The upper bound is quite large due to large second order derivatives of $V(x)$ and large time span T.

Then we run Algorithm 2 to get the contraction rate of $\hat{P}^T$ for $T = 50$. The number of initial values (x_i, y_i) is 20000. We run 1000 pairs of trajectories from each initial points to get $\mathbb{P}[\tau_c > T]$. This gives 20000 numbers,

$$r_i = \frac{\mathbb{P}_{x_i,y_i}[\tau_c > T]}{d(x_i, y_i)} \geq \frac{\mathrm{d}_w(\delta_{x_i}\hat{P}^T, \delta_{y_i}\hat{P}^T)}{d(x_i, y_i)}\,.$$

Then we use GPD to fit $\{1/(1 - r_i)\}_{i=1}^{20000}$. Similar to the previous example, GPD parameter fitting gives $\alpha_\Omega = 0.2157$. See Figure 2, top left, for the fitting result. In addition, because

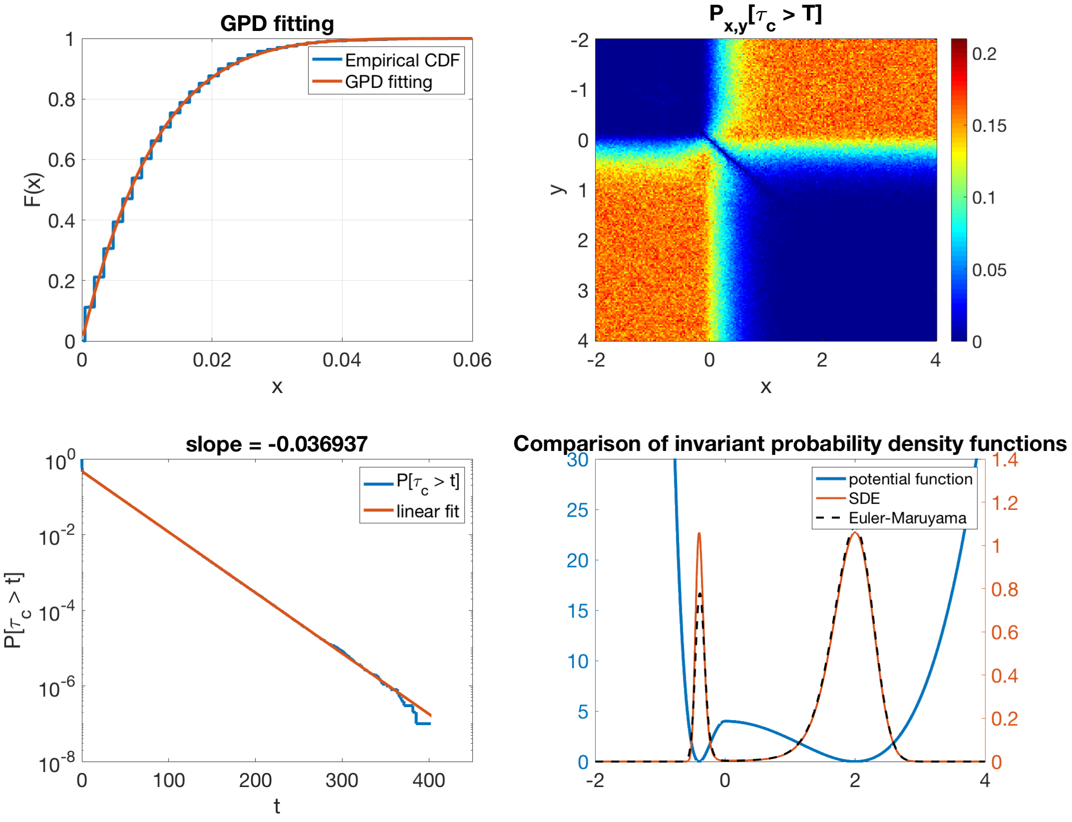

Figure 2. *Top left: Fitting GPD with $v_i = 1/(1 - r_i)$. The fitting result is compared with the empirical cumulative distribution function. Top right: Heat map of contraction rate r_i for initial pairs of points on a grid that covers $\Omega \times \Omega$. Bottom left: Exponential tail of $\mathbb{P}[\tau_c > t]$ versus t when initial values are uniformly sampled in $\Omega \times \Omega$. Bottom right: potential function $V(x)$ and a comparison of invariant probability density functions of $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}^h$.*

this is a 1D problem, we can plot the contraction rate r_i for each pair of (x_i, y_i) on a grid that covers $\Omega \times \Omega$. See Figure 2, top right, for a heat map of contraction rates from each pair of initial points. From Figure 2, right, we can see that the high value of $\mathbb{P}_{x_i,y_i}[\tau_c > T]$ is reached when one of the pairs (x, y) falls into the left part of the domain, where $V(x)$ has large derivatives. In addition, when (x, y) becomes even closer to line $\{x - y = 0\}$, $\mathbb{P}_{x_i,y_i}[\tau_c > T]$ drops dramatically to 0. This further confirms that $\hat{P}^T$ is contracting in 1-Wasserstein metric space for $T = 50$. Finally, we provide an exponential tail of $\mathbb{P}[\tau_c > t]$ for $\hat{\boldsymbol{X}}^h$, demonstrated in Figure 2, bottom left. The exponential tail is $\gamma = -0.036937$. Hence $e^{-\gamma T}$ gives 0.1577, which is smaller than α_Ω obtained above.

Combining all estimates above, we have an upper bound

$$\mathrm{d}_w(\pi, \hat{\pi}) \leq 0.2134\,. \tag{4.3}$$

If (3.7) is used instead, we have a rough estimate

$$\mathrm{d}_w(\pi, \hat{\pi}) \approx 0.19867\,.$$

Both results imply that two invariant probability measures may be very different from each other. This can be confirmed by using Monte Carlo simulation to compute the invariant probability measure of $\hat{\boldsymbol{X}}^h$. We run eight independent long trajectories of $\hat{X}^h_t$ up to 5×10^6 to compute its invariant probability density function. The result is compared with $u(x)$, the invariant probability density function of $\boldsymbol{X}$, in Figure 2, bottom right. We can see visible difference between these two probability density functions. The total variation difference between them is 0.05906. This is smaller than the bound predicted by (4.3), partially because we have to use the distance induced by $|x - y|^{0.45}$ to make r_i uniformly bounded from above. But our calculation still predicts an unusually large difference between two invariant probability measures.

We still owe readers a heuristic explanation of the phenomenon seen in Figure 2, bottom right. The probability density of the invariant probability measure of $\hat{\boldsymbol{X}}^h$ is much lower than that of $\boldsymbol{X}$ around the local minimum $x = -0.4$ because the potential function is asymmetric. As a result, when a trajectory of $\hat{\boldsymbol{X}}^h$ moves from $x = -0.4$ to $x = 0$, the Euler–Maruyama scheme tend to underestimate $-V'(x)$, which increases dramatically near $x = -0.4$. The effect of such underestimation is much weaker near the other local minimum $x = 2$, where the value of $|V'(x)|$ is significantly smaller. As a result, it is easier for the trajectory of $\hat{\boldsymbol{X}}^h$ to pass the separatrix $x = 0$ from left to right than from right to left. This causes the unbalanced invariant probability density function as seen in Figure 2, bottom right.

4.3. Degenerate diffusion.

Langevin dynamics have noise terms that only appear directly in the velocity equation and not on the position, leading to a Fokker–Planck equation with a degenerate, hypoelliptic diffusion. We consider a potential energy similar to the ring density equation, with SDE

$$\begin{aligned} dX_t &= V_t\,dt,\\ dV_t &= -\nabla U(X_t)\,dt - \gamma V_t\,dt + \sigma\,dW_t, \end{aligned} \tag{4.4}$$

where

$$U(X) = (X_1^2 + X_2^2 - 1)^2. \tag{4.5}$$

One trajectory of (4.4) is demonstrated in Figure 3, top left. The invariant measure satisfies $\rho(X, V) \propto \exp(-\beta(V^2 + U(X)))$ where $\beta = \frac{2\gamma}{\sigma^2}$.

Because of the degenerate noise term, we use a modified coupling algorithm involving three components (reflection coupling, synchronous coupling, and maximal coupling) which are based on the coupling in [16]. We consider two realizations $(X^{(1)}, V^{(1)})$ and $(X^{(2)}, V^{(2)})$ of the SDE and note that the difference process is contractive on the hyperplane $Q = X^{(1)} - X^{(2)} + \gamma^{-1}(V^{(1)} - V^{(2)}) = 0$ [16]. We employ reflection coupling when $\|Q\| > 0.08$, where the reflection tensor is given by $I - 2 * QQ^T/Q^2$. When $\|Q\| < 0.08$, we use synchronized coupling, where both processes use the same realization of the Brownian noise. The threshold of the switching coupling method (which is 0.08 in our computation) should be $O(\sqrt{h})$, which is the distance that $(\hat{X}_t^h, \hat{V}_t^h)$ jumps after one step. When the processes are sufficiently close, we attempt to couple using maximal coupling using two steps of the numerical integrator. Two steps of the Euler–Maruyama integrator with step size h gives

$$\begin{aligned} \hat{X}_{t+2h} &= \hat{X}_t + 2h\hat{V}_t - h^2\nabla U(\hat{X}_t) - \gamma h\hat{V}_t + \sigma h^{3/2}N_0, \\ \hat{V}_{t+2h} &= (1-\gamma h)^2\hat{V}_t - (1-\gamma h)h\nabla U(\hat{X}_t) - h\nabla U(\hat{X}_t + h\hat{V}_t) \\ &\quad + \sigma(1-\gamma h)h^{1/2}N_0 + \sigma h^{1/2}N_1, \end{aligned} \tag{4.6}$$

where N_0 and N_1 represent two independent and identically distributed normal variables. We sample from the above to compute the probability of the processes to couple after two steps. To improve the computational efficiency of the scheme we only test for coupling when the processes are close, specifically, when $|X^{(1)} - X^{(2)}| < 2.5\sigma h^{3/2}$ and $|V^{(1)} - V^{(2)}| < 2.5\sigma h^{1/2}$.

In this simulation, we choose $\sigma = 0.5$ and truncation time $T = 40$. The time step size is $h = 0.001$. Averaged from eight long trajectories with length 4×10^6, we find the error $d_w(\pi P^T, \pi\hat{P}^T) \le 0.0111313$. We choose the domain $\Omega = [-3\,,\,3]^2 \times [-6\,,6]^2$ for the coordinates (x_1, x_2, v_1, v_2). GPD fitting used 20000 pairs of initial values and 1000 trajectories for each pair of initial value giving $\alpha_\Omega = 0.5369$. See Figure 3, top right, for a comparison of the cumulative distribution function of the GPD fitting. This gives an upper bound

$$\mathrm{d}_w(\pi, \hat{\pi}) \le 0.02403\,.$$

It is not easy to numerically estimate the distance between π and $\hat{\pi}$ as they are probability measures in $\mathbb{R}^4$. Instead, we project them to the X-plane and compare the projected probability density functions. Note that the difference between two projected probability density functions is smaller than that of π and $\hat{\pi}$. In Figure 3, bottom left, we can see the difference between $P_x\hat{\pi}$ and $P_x\pi$, where P_x is the projection operator to the X-plane. The approximate numerical invariant measure $\hat{\pi}$ in Figure 3, bottom left, is obtained from 80 long trajectories, each of which is integrated up to $T = 10^7$. The probability density function of $\hat{\pi}$ is computed on a 512×512 grid.

Finally, as shown in Figure 3, bottom right, we compute the exponential tail of the coupling time to obtain the slope $\gamma = -0.025979$. Therefore, when $T = 40$, (3.7) gives a rougher estimate

$$\mathrm{d}_w(\pi, \hat{\pi}) \approx 0.01722\,.$$

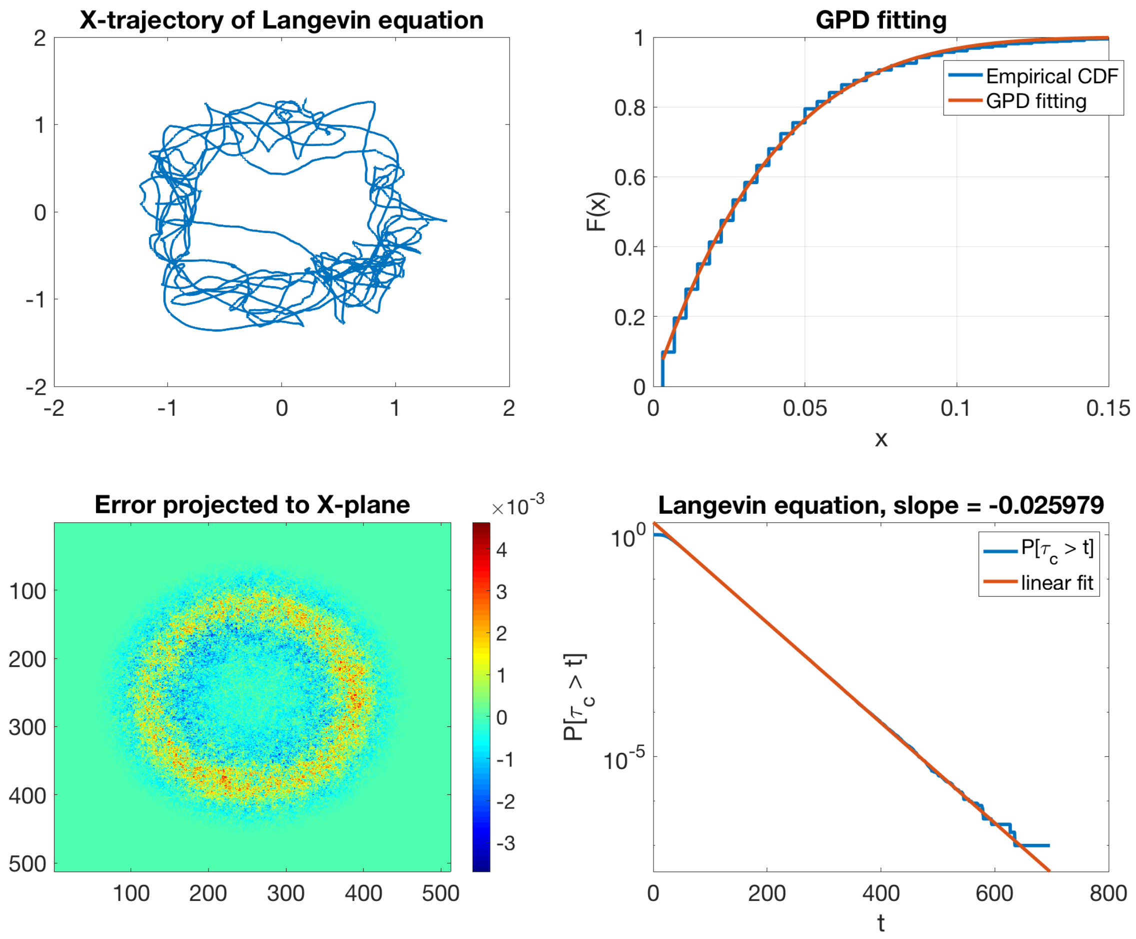

Figure 3. *Top left: A sample path of Langevin equation with length* 100 *(projected to X-plane). Top right: Fitting GPD with* $v_i = 1/(1 - r_i)$*. The fitting result is compared with the empirical cumulative distribution function. Bottom left: Difference between* $\hat{\pi}$ *and* π *projected to X-plane. Bottom right: Exponential tail of* $\mathbb{P}[\tau_c > t]$ *versus* t *when initial values are uniformly sampled in* $\Omega \times \Omega$*.*

4.4. Lorenz 96 model.

In this subsection we study a highly chaotic example. Consider equation

$$\begin{aligned}
\mathrm{d}X_t^1 &= (X_t^2 - X_t^{D-1})X_t^D - X_t^1 + F + \sigma \mathrm{d}W_t^{(1)}, \\
\mathrm{d}X_t^2 &= (X_t^3 - X_t^D)X_t^1 - X_t^2 + F + \sigma \mathrm{d}W_t^{(2)}, \\
&\vdots \\
\mathrm{d}X_t^i &= (X_t^{i+1} - X_t^{i-2})X_t^{i-1} - X_t^i + F + \sigma \mathrm{d}W_t^{(i)}, \quad i = 3, \ldots, D-1, \\
&\vdots \\
\mathrm{d}X_t^D &= (X_t^1 - X_t^{D-2})X_t^{D-1} - X_t^D + F + \sigma \mathrm{d}W_t^{(D)},
\end{aligned} \tag{4.7}$$

where the forcing term F is usually chosen to be 8. When $D = 4$, the system has a large periodic orbit. It demonstrates chaotic dynamics when $D \geq 5$ [27]. See Figure 4, top left, for the trajectory of the first three variables as an example.

In our simulations, we use the Euler–Mayurama scheme with step size $h = 0.0001$ to simulate numerical trajectories $\hat{X}^h_t$. Model parameters are $\sigma = 3$ and $F = 8$. The time span is chosen to be $T = 3$. When $D = 4$, in Algorithm 1, we run eight long trajectories with length 3×10^5 each to compare the difference between $\hat{X}^h_T$ and $\hat{X}^{2h}_T$. The simulation gives an upper bound

$$\mathrm{d}_w\left(\pi P^T, \pi \hat{P}^T\right) \leq 0.144864\,.$$

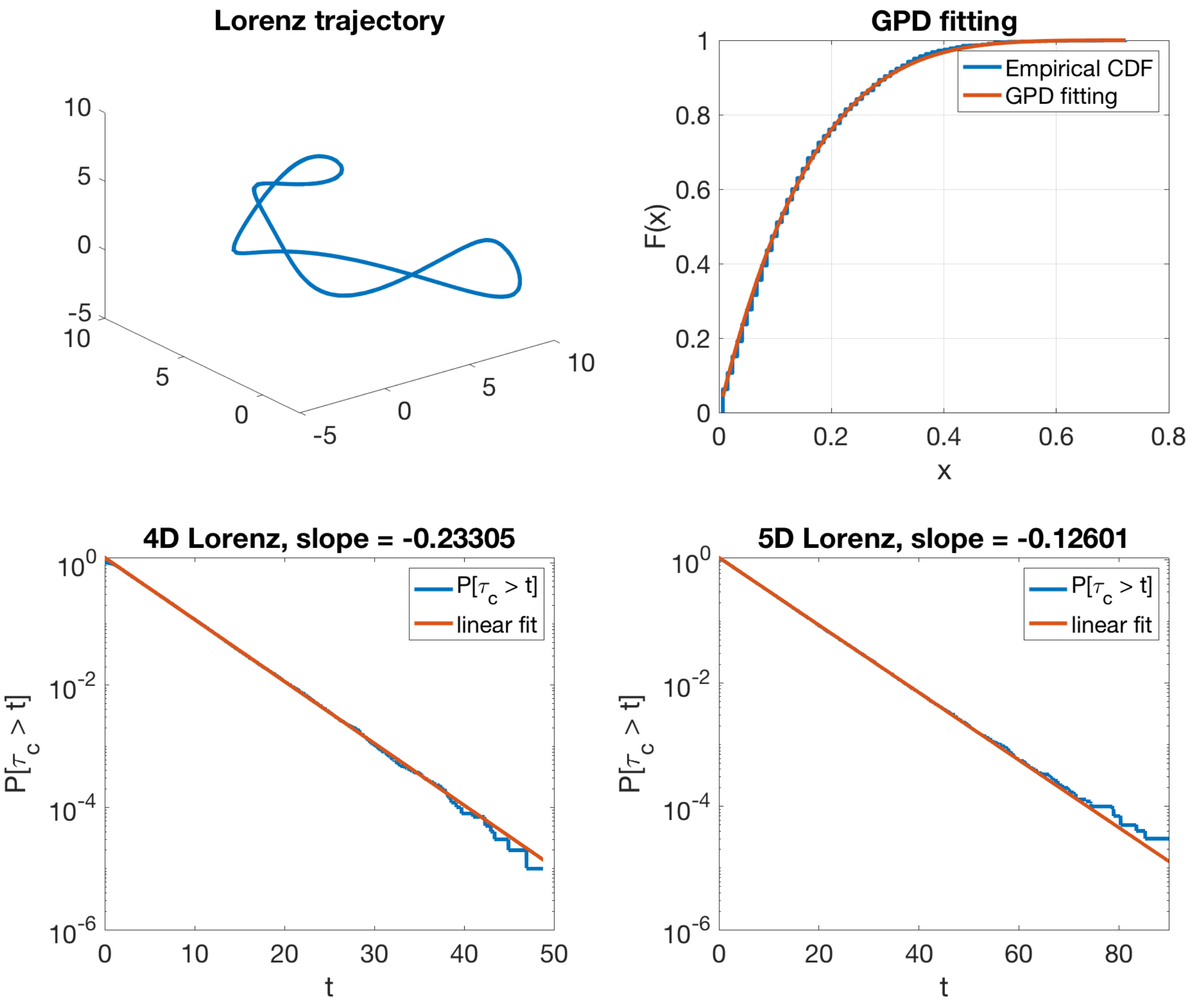

Figure 4. *Top left: A plot of the first three variables of the limit cycle. Top right: GPD fitting of* $\{1/(1 - r_i)\}_{i=1}^{20000}$ *and a comparison with the empirical cumulative distribution function. Bottom: Exponential tail of* $\mathbb{P}[\tau_c > t]$ *versus* t *when initial values are uniformly sampled in* $\Omega \times \Omega$*. Left:* 4*D Lorenz* 96 *system. Right:* 5*D Lorenz* 96 *system.*

Although we have chosen a small time step size h, this upper bound is still relatively large. The error gets even larger when $D = 5$ is used, because the deterministic dynamics is intensively chaotic. The output of Algorithm 1 is 0.11946 for $h = 0.00001$ and 0.431059 for $h = 0.0001$.

Then we run Algorithm 2 for $D = 4$ to get the contraction rate of $\hat{P}^T$ for $T = 3$. Ω is chosen to be the 4D box $[-16, 19]^4$ because when running Algorithm 1, no trajectory has ever been outside of this box. The number of initial values (x_i, y_i) is 20000. We run 1000 pairs of trajectories from each initial points to get $\mathbb{P}[\tau_c > T]$. This gives

$$r_i = \frac{\mathbb{P}_{x_i,y_i}[\tau_c > T]}{d(x_i, y_i)} \geq \frac{\mathrm{d}_w(\delta_{x_i}\hat{P}^T, \delta_{y_i}\hat{P}^T)}{d(x_i, y_i)}$$

for $i = 1, \dots, 20000$. The GPD fitting gives $\alpha_\Omega = 0.7302$. See Figure 4, top right, for the fitting result. Combine the output of two algorithms, we have the bound

$$\mathrm{d}_w(\pi, \hat{\pi}) \leq 0.5369\,.$$

When $D = 5$ and $h = 1 \times 10^{-5}$, the computational cost of Algorithm 2 becomes very high due to extremely small time step size. Instead, we compute exponential tails of the coupling time for $D = 4$ and $D = 5$ with $h = 0.0001$. The result is demonstrated in Figure 5, bottom. We can see that when $D = 5$, we have an exponential tail $\gamma = 0.12601$. Therefore, if $T = 3$ is unchanged, we have $(1 - e^{-\gamma T})^{-1} = 3.1767$. We conclude that when $D = 5$, the 1-Wasserstein distance between π and $\hat{\pi}$ is unacceptably large even if $h = 1 \times 10^{-5}$. This is mainly caused by very large finite time error. In order to approximate π effectively, high order approximation of (4.7) is necessary.

4.5. Stochastically coupled Fitzhugh–Nagumo oscillator with mean-field interaction.

We consider here a high dimensional example, a stochastically coupled Fitzhugh–Nagumo oscillator. The Fitzhugh–Nagumo model is a nonlinear model that models the periodic change of membrane potential of a spiking neuron under external stimulation. The model is a 2D system,

$$\begin{aligned} \mu \mathrm{d}u &= \left(u - \frac{1}{3}u^3 - v\right)\mathrm{d}t + \sqrt{\mu}\sigma \mathrm{d}W_t, \\ \mathrm{d}v &= (u + a)\mathrm{d}t + \sigma \mathrm{d}W_t\,, \end{aligned} \tag{4.8}$$

where u is the membrane potential, and v is a recovery variable.

When $a = 1.05$, the deterministic system admits a stable fixed point with a small basin of attraction [12]. A suitable random perturbation can drive this system away from the basin of attraction and trigger limit cycles intermittently.

In this section we consider N coupled equations (4.8) with both nearest-neighbor interaction and mean-field interaction. Let $v = \sqrt{\mu}v$ be the new recovery variable. We have

$$\begin{aligned} \mathrm{d}u_i &= \left(\frac{1}{\mu}u - \frac{1}{3\mu}u^3 - \frac{1}{\sqrt{\mu}}v + \frac{d_u}{\mu}(u_{i+1} + u_{i-1} - 2u_i) + \frac{w}{\mu}(\bar{u} - u_i)\right)\mathrm{d}t + \frac{\sigma}{\sqrt{\mu}}\mathrm{d}W_t^{(2i-1)}, \\ \mathrm{d}v_i &= \left(\frac{1}{\sqrt{\mu}}u + \frac{a}{\sqrt{\mu}}\right)\mathrm{d}t + \frac{\sigma}{\sqrt{\mu}}\mathrm{d}W_t^{(2i)} \end{aligned} \tag{4.9}$$

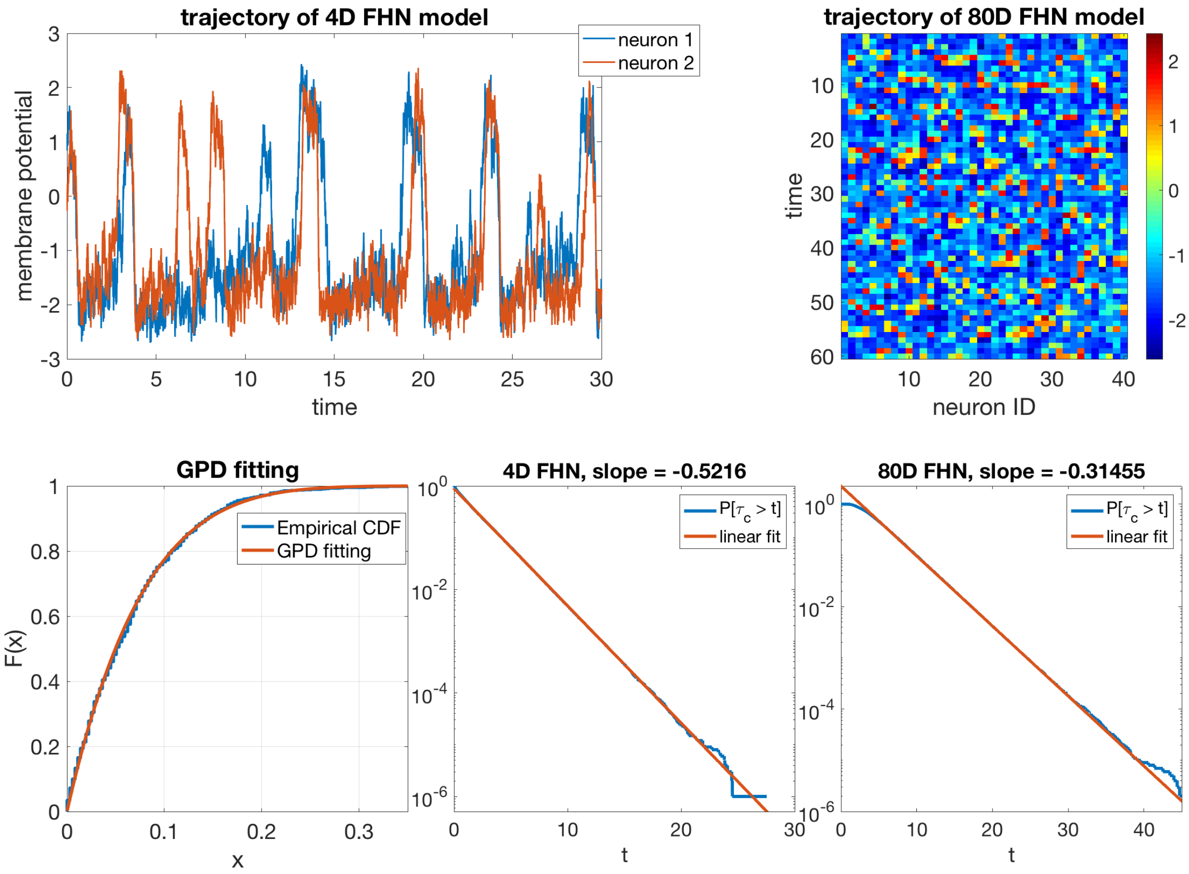

Figure 5. *Top: Dynamics of the Fitzhugh–Nagumo model. Left: Trajectory of membrane potential of two neurons evolving along* (4.9) *with* $N = 2$ *and other parameters specified in this paper. Right: Snapshots of membrane potential of* 40 *neurons evolving along* (4.9) *with* $N = 40$ *and other parameters specified in this paper. Bottom left: GPD fitting of* $\{1/(1-r_i)\}_{i=1}^{40000}$ *and a comparison with the empirical cumulative distribution function. Bottom middle: Exponential tail of* $\mathbb{P}[\tau_c > t]$ *versus* t *for* (4.9) *with* $N = 2$*. Bottom right: Exponential tail of* $\mathbb{P}[\tau_c > t]$ *versus* t *for* (4.9) *with* $N = 40$.

for $i = 1, \dots, N$, where $W_t^{(1)}, \dots, W_t^{(2N)}$ are independent Wiener processes, d_u is the neareast-neighbor coupling strength, w is the mean-field coupling strength, and

$$\bar{u} = \frac{1}{N}\sum_{i=1}^{N} u_i$$

is the mean membrane potential. In our simulations, we let $u_{-1} = u_N$ and $u_{N+1} = u_1$. In other words, N neurons are connected as a ring.

The parameters we choose are $d_u = 0.03$ and $w = 0.3$. In addition we have $\sigma = 0.6$. Activities of neurons are weakly coupled under this parameter set. See Figure 5, top left, for the dynamics of this system. In particular, from Figure 5, top right, we can see that nearest-neighbor neurons tend to spike together. However, the global dynamics is only weakly synchronized. The dimensions of the system in our study are chosen to be $N = 2$ and $N = 40$, corresponding to SDEs in $\mathbb{R}^4$ and $\mathbb{R}^{80}$. The numerical scheme in our simulation is an Euler–Maruyama scheme with $h = 0.0005$. The finite time span is $T = 3$ for both cases.

We first run Algorithms 1 and 2 for (4.9) with $N = 2$. In Algorithm 1, we run eight long trajectories up to 3×10^5. The simulation gives an upper bound

$$\mathrm{d}_w(\pi P^T, \pi \hat{P}^T) \leq 0.0105652$$

for $T = 3$. Then we run Algorithm 2 for $\Omega = [-6, 6]^{2N}$ to get the contraction rate of $\hat{P}^T$. The number of initial values is 40000. This gives 40000 coupling probabilities $r_1, \ldots, r_{40000}$. Fitting these numbers with GPD gives $\alpha_\Omega = 0.5271$. See the result in Figure 5, bottom left. Combining the output of two algorithms, we have

$$\mathrm{d}_w(\pi, \hat{\pi}) \leq 0.02234\,.$$

Hence the invariant probability measure simulated by running the Euler–Maruyama scheme is trustworthy in spite of the presence of slow-fast dynamics. In addition, we compute the tail of coupling time for $N = 2$, which is demonstrated in Figure 5, bottom middle. The exponential tail has a slope $\gamma = 0.5216$. Therefore, (3.7) gives an estimate

$$\mathrm{d}_w(\pi, \hat{\pi}) \approx 0.01336\,.$$

When $N = 40$, we still run Algorithm 1 with eight long trajectories up to time 3×10^5. This gives us an estimate

$$\mathrm{d}_w(\pi P^T, \pi \hat{P}^T) \leq 0.0443737$$

for $T = 3$. However, Algorithm 2 becomes expensive for $N = 40$. Instead we compute the exponential of coupling time to get a rough estimate. The exponential tail of coupling time is demonstrated in Figure 5, bottom right. We have an exponential tail with slope $\gamma = 0.31455$. Therefore, (3.7) gives a rough estimate

$$\mathrm{d}_w(\pi, \hat{\pi}) \approx 0.07265\,.$$

Therefore, we conclude that $\hat{\pi}$ is an acceptable approximation of π when $N = 40$.

5. Conclusion.

In this paper we provide a coupling-based approach to quantitatively estimate the distance between the invariant probability measure π of an SDE and that of its numerical scheme, denoted by $\hat{\pi}$. The key idea is that the distance $d(\pi, \hat{\pi})$ can be bounded by $\epsilon(1-\alpha)^{-1}$, where ϵ is the finite time truncation error over the time interval $[0, T]$, and α is the rate of contraction of $\hat{P}^T$, the time-T transition kernel of the numerical scheme for the SDE. The finite time truncation error comes from extrapolation analysis, and we use a coupling method to estimate α. Depending on the practical requirement, we provide one algorithm for computing a quantitative upper bound of $d(\pi, \hat{\pi})$ and an efficient algorithm for a "rough estimate" of $d(\pi, \hat{\pi})$. The performance of these two algorithms is tested with several numerical examples. Our approach can be extended to other stochastic processes, such as SDEs with random switching and applications related to Hamiltonian Monte Carlo [7, 36].

Essentially, the distance between two invariant probability measure studied in this paper is a sensitivity analysis problem. We study the robustness of π against a small change of the infinitesimal generator of an SDE caused by a time discretization. In practice, the small change of the infinitesimal generator does not have to come from a discretization. Therefore, our method is applicable to a large class of sensitivity analysis problems. For example, $\hat{\pi}$ could be the invariant probability measure of the original SDE subject to a small parameter change. The finite time truncation error can be replaced with the perturbation, and the speed of convergence can still be estimated by the coupling method.

This paper mainly estimates 1-Wasserstein distance between π and $\hat{\pi}$. However, the coupling method can also be used to estimate the other type of distances, such as the total variation distance. In addition, in many cases, we are actually more interested in the error of the expectation of a certain observable when integrating with respect to $\hat{\pi}$ versus π. We choose 1-Wasserstein distance mainly because it is more convenient to estimate the finite time truncation error in 1-Wasserstein distance. In fact, there is only a small literature about estimating finite time truncation error in the total variation norm. The difficulty of estimating finite time truncation error in total variation distance is partially solved if the grid-based SDE solver introduced in [8] is used. It is much easier to count samples on grids than in continuous state space. In fact, we find that this grid-based SDE solver is more compatible with both the Fokker–Planck solver in [32, 13] and the sample quality checking algorithm studied in this paper. In the future, we will write a separate paper to discuss the application of this sample quality checking algorithm to this grid-based SDE solver.

REFERENCES

[1] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, *An introduction to mcmc for machine learning*, Machine Learning, 50 (2003), pp. 5–43.

[2] D. Bakry and M. Émery, *Diffusions hypercontractives*, in Séminaire de Probabilités XIX 1983/84, Springer, New York, 1985, pp. 177–206.

[3] T. G. Bali, *The generalized extreme value distribution*, Econom. Lett., 79 (2003), pp. 423–427.

[4] A. Balkema and L. de Haan, *Residual life time at great age*, Ann. Probab., 2 (1974), pp. 792–804.

[5] V. Bally and D. Talay, *The law of the euler scheme for stochastic differential equations*, Probab. Theory Related Fields, 104 (1996), pp. 43–60.

[6] V. Bally and D. Talay, *The law of the Euler scheme for stochastic differential equations:* II. *Convergence rate of the density*, Monte Carlo Methods Appl., 2 (1996), pp. 93–128.

[7] N. Bou-Rabee, A. Eberle, and R. Zimmer, *Coupling and Convergence for Hamiltonian Monte Carlo*, preprint, https://arxiv.org/abs/1805.00452, 2018.

[8] N. Bou-Rabee and E. Vanden-Eijnden, *Continuous-Time Random Walks for the Numerical Solution of Stochastic Differential Equations*, Mem. Amer. Math. Soc. 256, AMS, Providence, RI, 2018.

[9] E. Castillo and A. S. Hadi, *Fitting the generalized pareto distribution to data*, J. Amer. Statist. Assoc., 92 (1997), pp. 1609–1620.

[10] B. Charbonneau, Y. Svyrydov, and P. F. Tupper, *Weak convergence in the Prokhorov metric of methods for stochastic differential equations*, IMA J. Numer. Anal., 30 (2010), pp. 579–594.

[11] C. Chen, J. Hong, and X. Wang, *Approximation of invariant measure for damped stochastic nonlinear schrödinger equation via an ergodic numerical scheme*, Potent. Anal., 46 (2017), pp. 323–367.

[12] N. Chen, A. J. Majda, and X. T. Tong, *Spatial Localization for Nonlinear Dynamical Stochastic Models for Excitable Media*, preprint, https://arxiv.org/abs/1901.07318, 2019.

[13] M. Dobson, Y. Li, and J. Zhai, *An Efficient Data-Driven Solver for Fokker-Planck Equations: Algorithm and Analysis*, https://arxiv.org/abs/1906.02600, 2019.

[14] C. R. Doering, K. V. Sargsyan, and P. Smereka, *A numerical method for some stochastic differential equations with multiplicative noise*, Phys. Lett. A, 344 (2005), pp. 149–155.

[15] A. Eberle, *Reflection coupling and wasserstein contractivity without convexity*, C. R. Math., 349 (2011), pp. 1101–1104.

[16] A. Eberle, A. Guillin, and R. Zimmer, *Couplings and quantitative contraction rates for langevin dynamics*, Ann. Probab., 47 (2019), pp. 1982–2010.

[17] M. I. Freidlin and A. D. Wentzell, *Random perturbations*, in Random Perturbations of Dynamical Systems, Springer, New York, 1998, pp. 15–43.

[18] M. Gelbrich and S. T. Rachev, *Discretization for stochastic differential equations, Ip Wasserstein metrics, and econometrical models*, in Distributions with Fixed Marginals and Related Topics, Lecture Notes Monograph Ser. 28, Institute of Mathematical Statistics, 1996, pp. 97–119.

[19] M. Hairer, *Convergence of Markov Processes*, Lecture Notes, University of Warwick, 2010, Available at http://hairer.org/notes/Convergence.pdf.

[20] M. Hairer and J. C. Mattingly, *Yet another look at Harris ergodic theorem for Markov chains*, in Seminar on Stochastic Analysis, Random Fields and Applications VI, Progr. Probab. 63, Springer, 2011, pp. 109–117.

[21] E. P. Hsu and K.-T. Sturm, *Maximal coupling of Euclidean Brownian motions*, Commun. Math. Stat., 1 (2013), pp. 93–104.

[22] W. Huang, M. Ji, Z. Liu, and Y. Yi, *Steady states of Fokker-Planck equations:* I. *Existence*, J. Dynam. Differential Equations, 27 (2015), pp. 721–742.

[23] W. Huang, M. Ji, Z. Liu, and Y. Yi, *Concentration and limit behaviors of stationary measures*, Phys. D, 369 (2018), pp. 1–17.

[24] P. E. Jacob, J. O'Leary, and Y. F. Atchadé, *Unbiased Markov Chain Monte Carlo with Couplings*, preprint, https://arxiv.org/abs/1708.03625, 2017.

[25] J. E. Johndrow and J. C. Mattingly, *Error Bounds for Approximations of Markov Chains Used in Bayesian Sampling*, preprint, https://arxiv.org/abs/1711.05382 2017.

[26] V. E. Johnson, *A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms*, J. Amer. Statist. Assoc., 93 (1998), pp. 238–248.

[27] A. Karimi and M. R. Paul, *Extensive chaos in the Lorenz*-96 *model*, Chaos, 20 (2010), 043105.

[28] R. Khasminskii, *Stochastic Stability of Differential Equations*, Stoch. Model. Appl. Probab. 66, Springer, New York, 2011.

[29] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Stoch. Model. Appl. Probab. 23, Springer, New York, 2013.

[30] B. Leimkuhler and C. Matthews, *Rational Construction of Stochastic Numerical Methods for Molecular Sampling*, Appl. Math. Res. Express AMRX, 2013 (2012), pp. 34–56.

[31] T. Lelievre and G. Stoltz, *Partial differential equations and stochastic methods in molecular dynamics*, Acta Numer., 25 (2016), pp. 681–880.

[32] Y. Li, *A data-driven method for the steady state of randomly perturbed dynamics*, Commun. Math. Sci., 17 (2019), pp. 1045–1059.

[33] Y. Li and Y. Yi, *Systematic measures of biological networks* I: *Invariant measures and entropy*, Commun. Pure Appl. Math., 69 (2016), pp. 1777–1811.

[34] T. Lindvall, *Lectures on the Coupling Method*, Courier Corporation, North Chelmsford, MA, 2002.

[35] T. Lindvall and L. C. G. Rogers, *Coupling of multidimensional diffusions by reflection*, Ann. Probab., 14 (1986), pp. 860–872.

[36] X. Mao and C. Yuan, *Stochastic Differential Equations with Markovian Switching*, Imperial College Press, London, 2006.

[37] J. C. Mattingly, A. M. Stuart, and D. J. Higham, *Ergodicity for SDES and approximations: Locally Lipschitz vector fields and degenerate noise*, Stoch. Process. Appl., 101 (2002), pp. 185–232.

[38] J. C. Mattingly, A. M. Stuart, and M. V. Tretyakov, *Convergence of numerical time-averaging and stationary measures via Poisson equations*, SIAM J. Numer. Anal., 48 (2010), pp. 552–577.

[39] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer, New York, 2012.

[40] G. N. Milstein and M. V. Tretyakov, *Computing ergodic limits for Langevin equations*, Phys. D, 229 (2007), pp. 81–95.

[41] A. Y. Mitrophanov, *Sensitivity and convergence of uniformly ergodic Markov chains*, J. Appl. Probab., 42 (2005), pp. 1003–1014.

[42] C. Mufa, *Estimation of spectral gap for Markov chains*, Acta Math. Sin. New Ser., 12 (1996), pp. 337–360.

[43] S. L. Nguyen and G. Yin, *Pathwise convergence rate for numerical solutions of stochastic differential equations*, IMA J. Numer. Anal., 32 (2012), pp. 701–723.

[44] J. Pickands III, *Statistical inference using extreme order statistics*, Ann. Statist., 3 (1975), pp. 119–131.

[45] G. O. Roberts, and R. L. Tweedie, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, 2 (1996), pp. 341–363.

[46] D. Rudolf, and N. Schweizer, *Perturbation theory for Markov chains via Wasserstein distance*, Bernoulli, 24 (2018), pp. 2610–2639.

[47] D. Talay, *Second-order discretization schemes of stochastic differential systems for the computation of the invariant law*, Stochastics, 29 (1990), pp. 13–36.

[48] D. Talay and L. Tubaro, *Expansion of the global error for numerical schemes solving stochastic differential equations*, Stoch. Anal. Appl., 8 (1990), pp. 483–509.